AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 11(8): 25744–25767.
DOI: 10.3934/math.20261031
Received: 30 June 2026
Revised: 21 July 2026
Accepted: 06 August 2026
Published: 19 August 2026

Research article

A deep learning framework for causal estimation using the cascade additive noise model: An optimization perspective

Sohail Ahmad[1], **Mona F. ElWakeel**[2,*], **Moiz Qureshi**[3,4], **Hasnain Iftikhar**[4,5,*] **and Paulo Canas Rodrigues**[6,7]

[1] Discipline of Statistics, The University of Agriculture Swat, Mingora 19120, Pakistan

[2] Department of Finance and Banking, College of Business Administration, Dar Al Uloom University, Riyadh 13314, Saudi Arabia

[3] Department of Statistics, University of Sindh, Jamshoro 76080, Pakistan

[4] Department of Statistics, Quaid-i-Azam University, Islamabad 45320, Pakistan

[5] Department of Statistics, University of Peshawar, Peshawar 25120, Pakistan

[6] Department of Statistics, Federal University of Bahia, Salvador 40170-110, Brazil

[7] Department of Business Management, University of Pretoria, Pretoria 0002, South Africa

* **Correspondence:** Email: hasnain@stat.qau.edu.pk, m.alwakel@dau.edu.sa.

Abstract: Causal inference in nonlinear, high-dimensional systems remains challenging due to complex dependencies and unobserved confounders. To address latent variables and indirect effects, a variational autoencoder (VAE) was incorporated into the cascade additive noise model (CANM). However, the CANM-VAE and its extension suffer from limitations in modeling long-range dependencies in high-dimensional settings because confounding variables are handled only at the latent-variable stage. To overcome these limitations, we developed the transformer-based variational CANM (TV-CANM), which leverages self-attention mechanisms to effectively capture complex nonlinear relationships and long-range dependencies. We trained the TV-CANM on three arrhythmia benchmark datasets, applied the test likelihood to the causal direction, and used the mean squared error (MSE) and root mean squared error (RMSE) to measure performance. Findings indicate that TV-CANM is significantly more effective than the original CANM at identifying causal directions consistent with domain knowledge and substantially reducing error rates. The proposed framework provides reliable causal knowledge that can support optimization-based decision-making and operational planning in complex data-driven systems. Moreover, the TV-CANM is found to be sound in nonlinear, high-dimensional causal discovery, compared with five traditional algorithms: additive noise model (ANM), information-geometric causal inference (IGCI), causal additive model (CAM), post-nonlinear model (PNL), and a loss-based algorithm.

Keywords: causal inference; cascade additive noise model; variational autoencoder; transformer; self-attention mechanism; high-dimensional data; optimization; operations research applications
Mathematics Subject Classification: 62H22, 68T07, 68T09, 90C26

1. Introduction

Causal inference establishes cause-and-effect relationships among variables, uncovering causality beyond observational relationships. One of the crucial problems in causal discovery is to identify the direction of causation, that is, whether variable X affects Y, or the variables are confounded by latent variables. This issue occurs in a variety of fields, such as medicine [1], biology [2], and the social sciences [3], in which resolving causal direction clarifies complex relations [4, 5].

Modern causal-discovery algorithms study data structures based on clear assumptions of variable relationships. As an example, in a biological context, we have the variables X (lifestyle factors such as diet and exercise) and Y (cardiovascular health). Causal relationship: It is possible that the causal relationship between the unobservable intermediate variable, C (e.g., inflammation markers), lies between the two variables (i.e., Figure 1). This is where C passes on the influence of X onto Y though it is not visible. These situations highlight the importance of modeling latent variables to derive accurate causal pathways, especially in complex biological systems.

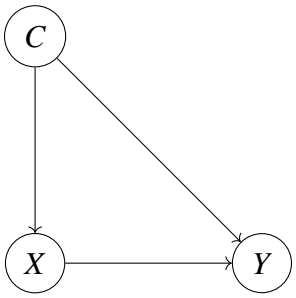


Figure 1. Causal inference with input variable (X), output variable (Y), and hidden variable (C).

Causal inference methods are generally developed in two phases: first, preliminary evaluation of cause-and-effect relationships, and second, use of mediator variables (intermediate variables) to eliminate bias [11]. The most common methods of dealing with intermediate variables are: randomized control trials (RCTs) [6], propensity score matching [7], instrumental variables [8], regression discontinuity [9], mediation analysis [10], marginal structural equation modeling [12], Bayesian networks [13], synthetic control methods [14], and causal-risk method [15]. In the meantime, causal discovery has also evolved and moved past the traditional statistical approaches to the data-driven methods of constraint-based like the Peter–Clark (PC) algorithm [17] approaches to causal discovery, 2001 score-based (like greedy equivalence search (GES) [17]), and the causal order discovery (CONOD) algorithm [16, 18, 19].

However, these methods tend to be restricted in the presence of nonlinear relationships, cyclical relationships, and high-dimensional data. Such complications are overcome by functional causal models (FCMs) [20], which explicitly model relationships as functional equations and are especially effective in

complex, nonlinear causal structures. One of the most notable subclasses of FCMs is the additive noise model (ANM) [21], which uses independence between the terms that cause something and the noise to create causal asymmetry. Another notable ANM implementation is the linear non-gaussian acyclic model (LiNGAM) [22] algorithm for linear non-Gaussian acyclic models with time-series data and latent variables. Although ANMs are interpretable and highly generalizable, they impose demanding assumptions, such as non-Gaussian noise and the absence of confounders in the hidden variables. The causal direction is normally established by estimating the functional model and evaluating independence between the proposed causes and the residual noise [21, 23, 24].

Continuous optimization approaches overcome FCM limitations by redefining causal discovery as differentiable learning. Scalable graph inference methods such as non-combinatorial optimization via trace exponential and augmented lagRangian for structure learning (NOTEARS) [25], directed acyclic graph–deep neural network (DAG-DNN) [26], and efficient neural causal optimization (ENCO) [27] are based on deep learning. In addition to improving efficiency in high-dimensional spaces, these techniques are sensitive to initialization and require tuning. Such hybrid strategies that involve domain knowledge as max-min hill-climbing (MMHC) [28] and joint causal inference JCI [29] are more reliable, but have separate scalability limitations [30]. Flexible alternatives are provided by latent representations based on data from generative deep learning models. This bypasses the strong assumptions of linearity and initialization when learning complex distributions in the latent space of variational autoencoders (VAEs) [31–33], allowing the uncovering of high-dimensional causal patterns. In this regard, adaptive operator selection with deep Q-network (CANM-VAE) was proposed in [34] for multi-objective optimization, and the variational autoencoder (VAE) was incorporated into the additive noise framework to learn the latent representation, which enhanced the discrimination capability of the causal mechanism in high-dimensional environments with complex dependency relationships [35].

On the basis of this research, [36] proposed a transformer variational autoencoder model for individual treatment effect estimation, [37] suggested a structural equation model with machine learning techniques to improve causal effect estimation, and [38] proposed an individual treatment effect estimation model called the transformer variational causal autoencoder (TV-CCANM). Despite these advances, TV-CCANM models confound variables implicitly at the latent stage, which limits interpretability and makes the underlying causal mechanisms difficult to analyze.

Motivated by these limitations, we propose a transformer-based variational autoencoder within the cascade additive noise model (TV-CANM). The proposed model extends both the CANM-VAE and TV-CCANM by leveraging self-attention mechanisms to explicitly capture complex interactions and long-range dependencies in high-dimensional data, while preserving a clearer and more interpretable causal structure. This extension provides a more accurate and robust framework for dependency modeling and causal structure identification in complex systems. The CANM-VAE extends the cascade additive noise model by incorporating variational inference to improve latent representation learning for nonlinear causal discovery. Building upon this idea, the TV-CCANM further integrates transformer-based variational inference to model confounding cascade structures through self-attention mechanisms. In contrast, the proposed TV-CANM is specifically designed for causal direction identification within the cascade additive noise model by replacing the conventional VAE encoder with a transformer-based encoder. The model uses a multi-head self-attention mechanism to learn complex nonlinear interactions while maintaining the original cascade additive noise model and capturing long-range dependencies. Consequently, the TV-CANM boosts latent representation learning and causal direction estimation

without introducing the confounding cascade structure in the case of the TV-CCANM. Thus, our major contributions are:

- The TV-CANM integrates a transformer-based variational autoencoder into the cascade additive noise model for nonlinear causal estimation.
- The proposed self-attention mechanism captures long-range dependencies that cannot be effectively modeled by conventional CANM.
- The proposed framework improves causal direction identification in high-dimensional settings while maintaining model interpretability.
- Extensive experiments on benchmark arrhythmia datasets demonstrate superior performance over CANM and existing causal discovery approaches.

2. Preliminaries

In this section, we present some notation and assumptions describing the structural modeling technique used in the transformer-based cascadeadditive noise model (CANM).

2.1. Problem formulation and notation

Let $X, Y \in \mathbb{R}$ be two observed real-valued random variables. Let $\eta = \{\eta_1, \ldots, \eta_T\} \subset \mathbb{R}^T$ denote a sequence of unobserved intermediate variables that form a cascade between X and Y (see Figure 2). In the system, every variable is produced by a deterministic nonlinear function of its parent variables, perturbed by an independent additive noise term.

2.1.1. Definitions

We present some common definitions.

Definition 1: Additive Noise Model

An additive noise model (ANM) is a structural equation of the form

$$Y = f(X) + \epsilon, \tag{2.1}$$

where f is a deterministic function and ϵ is a noise factor that is statistically independent of X. ANMs provide identifiability of causal direction under moderate confounding and are commonly used in nonlinear causal modeling.

Definition 2: Cascade structure

A cascade structure is a series of latent variables $\eta_1, \eta_2, \ldots, \eta_T$, each of which is influenced by its previous generations; the final variable η_T shapes the result Y. This sequence illustrates the indirect transfer of causal effect from X to Y as follows:

$$X \to \eta_1 \to \eta_2 \to \cdots \to \eta_T \to Y.$$

Definition 3: Cascade Additive Noise Model

Defined by the following equations, a nonlinear structural model including observable variables X and Y, and intermediate variables $\eta_1, \ldots, \eta_T$, is a cascade additive noise model (CANM).

$$\begin{aligned} \eta_1 &= f_1(X) + \xi_1, \\ \eta_t &= f_t(\eta_{pa(t)}) + \xi_t, \quad t = 2, \ldots, T, \\ Y &= f_{T+1}(\eta_{pa(Y)}) + \epsilon_Y, \end{aligned} \tag{2.2}$$

where $f_X, f_1, \ldots, f_{T+1}$ are deterministic nonlinear functions, $\epsilon_X, \epsilon_Y, \xi_1, \ldots, \xi_T$ are mutually independent noise terms, and $pa(t) \subseteq \{\eta_1, \ldots, \eta_{t-1}\}$, $pa(Y) \subseteq \{\eta_1, \ldots, \eta_T\}$ denote the parent sets of η_t and Y, respectively.

2.1.2. Assumptions

The following assumptions are needed:

- **Assumption 1: Additive noise**
  Each variable in the system is created as a nonlinear deterministic function of its parent variables with an additive noise term.

  $$X = f_X + \epsilon_X, \quad Y = f_{T+1}(\eta_{pa(Y)}) + \epsilon_Y,$$

  where f_X and f_{T+1} are deterministic nonlinear functions, and ϵ_X and ϵ_Y are independent noise terms.
- **Assumption 2: Noise independence**
  All noise components $\epsilon_X, \epsilon_Y, \xi_1, \ldots, \xi_T$ are mutually independent and are also independent of their related functional inputs. Mathematically:

  $$\epsilon_Y \perp\!\!\!\perp \{\eta_1, \ldots, \eta_T\}, \quad \xi_t \perp\!\!\!\perp \{\eta_{pa(t)}\}, \quad \forall t \in \{1, \ldots, T\}.$$

  In practice, these assumptions are adopted from the standard additive noise model framework. The independence between residual noise and predictor variables is expected after the transformer encoder captures the nonlinear causal mechanism. Therefore, the residual component represents unexplained random variation instead of systematic dependence.
- **Assumption 3: Causal pathway via intermediate variables**
  A series of intermediary variables $\eta_1, \ldots, \eta_T$ influences the effects of X on Y, creating an indirect causal path. The structural equations for the intermediate variables are as follows:

  $$\eta_1 = f_1(X) + \xi_1, \quad \eta_t = f_t(\eta_{pa(t)}) + \xi_t, \quad t = 2, \ldots, T.$$

  The cascade of intermediate variables represents how X effects Y through $\eta_1, \ldots, \eta_T$.

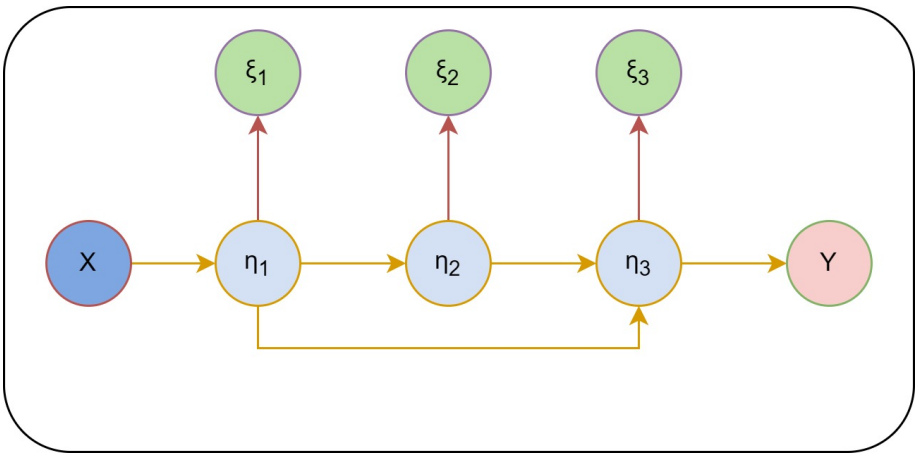


Figure 2. Causal inference with treatment variable (X), outcome (Y), unobserved intermediate variable (η), and noise term (ξ).

2.2. *Transformer model architecture*

The transformer model architecture was developed by the authors of [39] to facilitate direct parallel processing and to discover complex relationships in sequential data [40]. This model consists of two components, namely, the encoder and the decoder. The encoder consists of two identical blocks, each containing a multi-head attention (MHA) layer and a feedforward network (FFN) layer. To preserve the sequence order, the input data (X) is first converted to a vector space, and then positional encoding is applied, transforming it into a high-dimensional vector space that encodes important features. This embedding is processed by the first MHA layer, which uses heads (h). The linear versions of the queries (Q), keys (K), and values (V) matrices are learned and fed to each of the heads. These heads operate simultaneously to generate (h) outputs, which are later combined to make the final output. It is then followed by a normalization step and a residual connection, which involves feeding the attention layer's output back into the input. The second layer is the FFN, denoted by Eq (2.3), and performs two linear transformations with a rectified linear unit (ReLU) [41] in the middle. The FFN carries out the initial linear transformation, the activation function, and then a second linear transformation. This is followed by layer normalization. There is also a second residual connection that incorporates both the FFN input and its output.

$$FFN(x) = ReLU\Big(W_1X + b_1\Big)\cdot W_2 + b_2. \tag{2.3}$$

The weight and bias values of the encoder are (W_1 and W_2) and (b_1 and b_2) to implement the transformations of every input sequence. The encoder performs sequence interpretation and contextual understanding by implementing this process twice through its two blocks. The transformer encoder maintains its structure as shown in Figure 4.

2.2.1. Multi-head attention mechanism

The self-attention mechanism is a key operational element in transformer models and has many applications in machine translation and text generation systems. The mechanism develops three vector groups for sequences through the combined functions of queries (Q), keys (K), and values (V). The

vectors derive from learned linear transformations.

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V.$$

The sequence is labeled as X, while the weight matrices are identified as W_Q, W_K, and W_V. To determine the attention scores, which reflect how different segments of the sequence relate to each other, the dot product of the queries and keys is computed. This outcome is then adjusted by dividing by the square root of the keys' dimension (d_k), followed by the application of the softmax function.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.4}$$

The gradient stability varies with the scaling factor $\sqrt{d_k}$ during the training process, so that the value of the dot product remains within a realistic range.

Once the softmax function is applied, it yields a probability distribution that reflects the dominant characteristics of the sequence. The model considers the crucial components of the sequence at specific points and will calculate the sum of the values at the input points weighted by the key points. The transformer uses different attention heads to look at different aspects of the incoming data. Each head then has an independent self-attention mechanism, followed by a combination mechanism that applies a linear transformation to the output of that head.

$$\text{MultiHead}\Big(Q, K, V\Big) = \text{Concat}\Big(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\Big) \cdot W_O, \tag{2.5}$$

where each head i is calculated by the following:

$$\text{head}_i = \text{Attention}\Big(QW_{Q_i}, KW_{K_i}, VW_{V_i}\Big). \tag{2.6}$$

In the model presented in Figure 3 below, multiple sets of weights are provided to analyze different components of the sequence and help the model recognize complex sequence properties.

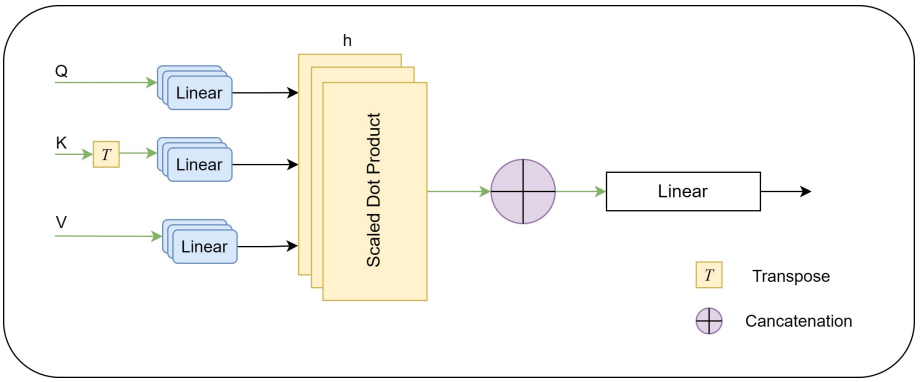


Figure 3. Multi-head attention mechanism.

The self-attention algorithm enhances the selection of features for causal inference. Dynamic importance evaluation is used to detect significant features that connect relationships between treatment and outcome effects. In this way, the model eliminates irrelevant factors and generates more precise and understandable results about causal effects.

The transformer encoder uses self-attention to learn from a dynamic model of all the input features, whereas a typical VAE is based on fully connected neural networks for feature extraction. This makes the latent representation very powerful: it not only captures local nonlinear interactions but also the relationships between features found across the whole process. This better representation results in better reconstruction and identification of causal direction.

3. The proposed method

The proposed method develops a transformer-based cascadeadditive noise model (TV-CANM) for determining causal relationships in complex datasets. Our discussion begins by presenting the transformer-CANM as a framework that improves on the standard CANM by integrating self-attention to enable longer-range dependency recognition. The marginal distribution of the TV-CANM is developed next before the model is approximated through a transformer-based encoder decoder system. Various components of the generative model are explained in detail, along with a method for refining representations using variational inference. The TV-CANM framework has an identifiability theorem that provides theoretical support. A complete algorithm uses the test likelihood together with the reconstruction error to establish the causal direction.

3.1. The transformer cascade additive noise model

The TV-CANM represents a causal system as a framework for situations in which multiple hidden variables (η_t) form a chain that affects the connection between X and Y. Under this model, there is a latent variable η_i that mediates the causal effect of X on Y. The model is defined as follows:

$$\begin{aligned} \eta_1 &= f_{1(\text{Transformer})}(X) + \xi_1, \\ \eta_t &= f_{t(\text{Transformer})}(\eta_{\text{pa}(t)}) + \xi_t, \quad t = 2, \ldots, T, \\ Y &= f_{T+1(\text{Transformer})}(\eta_{\text{pa}(Y)}) + \epsilon_y, \end{aligned} \tag{3.1}$$

where $\eta_{\text{pa}(t)}$ represents the parent variables of η_t in the causal chain. $f_{\text{Transformer}}(\cdot)$ denotes the transformer encoder, which maps the inputs to the next intermediate variable in the chain. $X, \xi_1, \xi_2, \ldots, \xi_T$ and ϵ_y are mutually independent noise terms.

3.1.1. Marginal distribution in the TV-CANM

The first step is to derive the marginal likelihood of the observed variables X and Y by integrating out the unobserved intermediate variables $\eta_1, \eta_2, \ldots, \eta_t$. We start with the joint probability:

$$p(X, Y, \eta_1, \eta_2, \ldots, \eta_t) = p(X)p(\eta_1|X)p(\eta_2|\eta_1) \cdots p(\eta_t|\eta_{t-1})p(Y|\eta_t). \tag{3.2}$$

Using the additive noise assumption for each relationship in the chain, we express the conditional distribution as:

$$\begin{aligned} p(\eta_1|X) &= p(\xi_1 = \eta_1 - f_{1(\text{Transformer})}(X)), \\ p(\eta_t|\eta_{t-1}) &= p(\xi_t = \eta_t - f_{t(\text{Transformer})}(\eta_{\text{pa}(t)})), \quad t = 2, \ldots, T, \\ p(Y|\eta_t) &= p(\epsilon_y = Y - f_{t+1(\text{Transformer})}(\eta_{\text{pa}(Y)})). \end{aligned} \tag{3.3}$$

To compute the marginal log-likelihood of the observed data X and Y, we integrate all intermediate variables. We set the data $\mathcal{D} = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{k}$, and let ϕ be the causal mechanism parameter to combine all independence relations. The following marginal distribution gives this:

$$\begin{aligned}
&\log \prod_{i=1}^{k} \int \left(x^{(i)}, y^{(i)}, \eta\right) d\eta \\
&= \log \prod_{i=1}^{k} \int p_\phi(x^{(i)}) p_\phi(y^{(i)}|\eta_{pa(y)}) \prod_{t=2}^{T} p_\phi(\eta_t|(\eta_{\mathrm{pa}(t)})) p_\phi(\eta_1|x^{(i)})\, d\eta, \\
&= \log \prod_{i=1}^{k} \int p_\phi(x^{(i)}) p_\phi(\epsilon_y^{(i)} = y^{(i)} - f_{(Transformer)}(x^{(i)}, \nu)) \prod_{t=1}^{T} p_\phi(\nu_t)\, d\nu, \\
&= \log \prod_{i=1}^{k} \int \left(x^{(i)}, \epsilon_y^{(i)}, \nu\right) d\nu.
\end{aligned} \tag{3.4}$$

Here, $p_\phi(x^{(i)})$ represents the prior for the input variable X. $p_\phi(\eta_t|(\eta_{\mathrm{pa}(t)}))$ models the conditional probability of η_t using the transformer encoder. $p_\phi(y^{(i)}|\eta_t)$ represents the likelihood of Y given the latent variable η_t, modeled using the transformer decoder. For the TV-CANM, we decompose the joint likelihood using the Markov property [17], ensuring that the latent variables are independent of the noise. We rewrite the transformation from the latent space η to the observed space Y using a function approximator within the VAE. The transformation captures the causal relation between X and Y, with the noise ξ accounted for, ensuring the identifiability of the causal direction. The variational inference of the VAE, which is used in (3.4), estimates the marginal log-likelihood, which helps determine the cause-and-effect relationship among the variables.

3.1.2. Variational solution of the TV-CANM

The variational method of the TV-CANM leverages transformers to process sequential information, with the goal of minimizing a variational lower bound to estimate latent variables. In the context of the TV-CANM, a variational-based technique is used because the latent variables are represented by high-dimensional entities and thus the marginal likelihood is complex. The algorithm exploits the latent noise independency to replace latent η elements of the form of the hidden ξ noise with the amortized inference distribution, where $q_\Theta(\xi|X, Y)$ represents the amortized inference distribution as follows: $q_\Theta(\xi|X, Y)$ is used to replace the hidden noise elements, $p_\phi(\xi|X, Y)$. This approach yields the greatest marginal log-likelihood lower-bound optimum with noise removal between the data of X and Y without altering the cause-and-effect relationships between the data.

The model performs data transformation by mapping temporal sequence inputs with their corresponding results into a latent dimension. Transformer structure is a learning process of the encoder as it learns the posterior distribution, $q_\Theta(\eta|X, Y)$, during the process of encoding. The latent variables η contain data from the input variables X and output variables Y. The latent space function captures all underlying structure found in the connections between input and output elements. The decoder performs data reconstruction after it derives this latent space by estimating the likelihood $p_\theta(X|\eta)$. The objective is to enhance the variational lower bound log-likelihood by optimizing its two constituents, which include reconstruction terms alongside Kullback-Leibler (KL) divergence

regularization terms. This can be formulated as:

$$\begin{aligned}
&\log \prod_{i=1}^{k} \int \left(x^{(i)}, \epsilon_y^{(i)}, \nu\right) d\nu \\
&= \sum_{i=1}^{k} E_{\nu \sim q_\Theta(\nu|x^{(i)}, x^{(i)})} \left[\log p_\phi(x^{(i)}, y^{(i)}|\nu)\right] - KL(q_\Theta(\nu|x^{(i)}, y^{(i)})||p_\phi(\nu|x^{(i)}, y^{(i)})) \\
&= \sum_{i=1}^{k} E_{\nu \sim q_\Theta(\nu|x^{(i)}, x^{(i)})} \left[\log \frac{p_\phi(x^{(i)}, \epsilon_y^{(i)}, \nu)}{q_\Theta(\nu|x^{(i)}, y^{(i)})}\right] - KL(q_\Theta(\nu|x^{(i)}, y^{(i)})||p_\phi(\nu|x^{(i)}, y^{(i)})) \\
&\leq \sum_{i=1}^{k} \mathcal{L}(\phi, \Theta, x^{(i)}, y^{(i)}).
\end{aligned} \tag{3.5}$$

The initial term promotes the precise reconstruction of the input $X^{(i)}$ using the latent variables η, while the second term represents the KL divergence, which serves to regularize the inferred posterior $q_\Theta(\nu|X^{(i)}, Y^{(i)})$ so that it remains close to a prior distribution $p_\phi(\nu)$, typically assumed to be a standard normal distribution.

The evidence lower bound (ELBO) and KL divergence equations can be expressed as:

$$\begin{aligned}
&\sum_{i=1}^{k} \mathcal{L}(\phi, \Theta, x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{k} \mathbb{E}_{\nu \sim q_\Theta(\nu|x^{(i)}, y^{(i)})} \left[-\log q_\Theta(\nu|x^{(i)}, y^{(i)}) + \log p_\phi(x^{(i)}, \epsilon_y^{(i)}, \nu)\right] \\
&= \sum_{i=1}^{k} \log p(|x^{(i)}) - \mathrm{KL}(q_\Theta(\nu|x^{(i)}, y^{(i)})||p_\phi(\nu)) \\
&\quad + \mathbb{E}_{\nu \sim q_\Theta(\nu|x^{(i)}, y^{(i)})} \left[\log p(\epsilon_y^{(i)} = y^{(i)} - f_{(Transformer)}(x^{(i)}, \nu, \phi))\right],
\end{aligned} \tag{3.6}$$

and

$$\mathrm{KL}(q_\Theta(\nu|x^{(i)}, y^{(i)})||p_\phi(\nu) = \frac{1}{2} \sum_{i=1}^{d_z} \left(1 + \log(\sigma_{\phi,i}^2) - \mu_{\phi,i}^2 - \sigma_{\phi,i}^2\right). \tag{3.7}$$

The detailed derivation is given in the supplementary file (Appendix). There, d_z is the dimensionality of the latent space, and $\mu_\phi(X, Y)$ and $\sigma_\phi(X, Y)$ are the mean and variance of the latent variables predicted by the transformer encoder.

The latent variables learned by the transformer-based variational autoencoder should not be interpreted as directly observed biological measurements. Instead, they represent compact latent representations that summarize unobserved physiological processes influencing the causal pathway. For example, in the relationship between age and body weight, these latent variables may capture the combined influence of metabolic activity, hormonal regulation, physical condition, and other unmeasured biological factors.

3.2. The TV-CANM model

The TV-CANM model is a transformer-based variational autoencoder designed to infer causal direction in nonlinear, high-dimensional datasets. It handles input pairs (X, Y), where X represents the cause and Y the effect, by encoding them into a latent variable ν. The analyzed variable joins X to form the reconstruction of Y. The transformer architecture of the encoder displays efficient capabilities for detecting long-distance relationships and a fully connected neural network design defines the decoder function. The architecture system appears in Figure 4.

Linear projection functions as the starting point of encoding where it transforms the input (X, Y) before ReLU activation occurs. The multi-layer transformer encoder operates on this representation in subsequent processing. The computation of the mean and log-variance of the latent variable occurs after the output processing stage:

$$h_\Theta(X, Y) = f_{\text{TransformerEncoder}}(\text{ReLU}(\text{Linear}([X, Y]))), \tag{3.8}$$

$$\mu_\Theta(X, Y) = \text{Linear}(h_\Theta(X, Y)), \quad \log \sigma^2_\Theta(X, Y) = \text{Linear}(h_\Theta(X, Y)). \tag{3.9}$$

To enable stochastic sampling during training while preserving differentiability, the reparameterization trick is used:

$$\nu = \mu_\Theta(X, Y) + \sigma_\Theta(X, Y) \cdot u, \quad \text{where } u \sim \mathcal{N}(0, I), \tag{3.10}$$

and $\sigma_\Theta(X, Y) = \exp\left(0.5 \cdot \log \sigma^2_\Theta(X, Y)\right)$.

A feedforward neural network contained within the decoder uses the concatenated vector $[X, \nu]$ to reconstruct the output Y:

$$\hat{Y} = f_{\text{Decoder}}([X, \nu]). \tag{3.11}$$

This network uses the function $f_{\text{Decoder}}(\cdot)$ as its fully connected component.

The model is trained by minimizing two loss terms: reconstruction error and KL divergence. Under the assumption of a Gaussian distribution with fixed standard deviation sdy, the reconstruction loss will estimate stability for numerical stability purposes:

$$\mathcal{L}_{\text{recon}} = -\sum \log \mathcal{N}(Y - \hat{Y}; 0, sdy^2), \tag{3.12}$$

and the KL divergence regularizes the latent space:

$$\mathcal{L}_{\text{KL}} = -0.5 \sum \left(1 + \log \sigma^2_\Theta(X, Y) - \mu^2_\Theta(X, Y) - \sigma^2_\Theta(X, Y)\right). \tag{3.13}$$

The total loss is given by:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{KL}}, \tag{3.14}$$

where β controls the trade-off between reconstruction fidelity and regularization.

After training, the model's ability to identify the correct causal direction is assessed using the test log-likelihood for both $X \to Y$ and $Y \to X$. The direction with the higher likelihood is selected as the inferred causal direction. Additionally, MSE and RMSE are used to compare model performance.

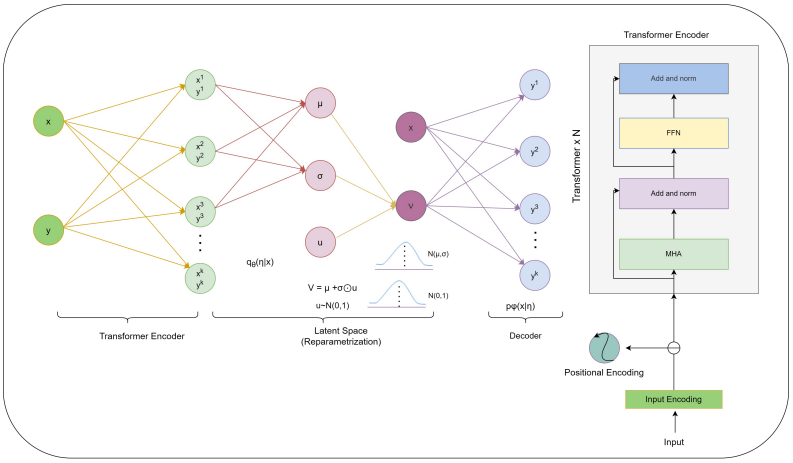


Figure 4. Transformer variational autoencoder for the TV-CANM.

Although the experimental evaluation considers selected physiological variables for interpretability, the proposed architecture is designed for high-dimensional applications. The transformer encoder processes all input dimensions simultaneously through multi-head self-attention, allowing efficient modeling of complex nonlinear dependencies without requiring handcrafted feature selection.

3.3. TV-CANM identifiability

The TV-CANM model is identifiable if the noise $\hat{\epsilon}$ is independent of both Y and $\hat{\xi}$. In other words, if the marginal likelihoods in both directions are equal, the causal direction $X \to Y$ can be uniquely determined.

Theorem 3.1. *Let $X \to Y$ follow a TV-CANM model, where a backward model also exists in the form:*

$$\begin{aligned} Y &= f_{Transformer}(X, \xi) + \epsilon, \quad X, \xi, \text{ and } \epsilon \text{ are independent} \Rightarrow \text{Forward Model} \\ X &= g_{Transformer}(Y, \hat{\xi}) + \hat{\epsilon}, \quad Y, \hat{\xi}, \text{ and } \hat{\epsilon} \text{ are independent} \Rightarrow \text{Backward Model.} \end{aligned} \tag{3.15}$$

Then, the noise distribution of the reverse direction $p_{\hat{\epsilon}}$ must be:

$$p_{\hat{\epsilon}}(\hat{\epsilon}) = \int e^{2\pi i \hat{\epsilon} \cdot v} \frac{\iint \rho(x) p(v) p_{\epsilon}(y - f(x, v)) e^{-2\pi i x \cdot v} \, dx \, dv}{p(y) \int p(\hat{v}) e^{-2\pi i g(y, \hat{v}) \cdot v} \, d\hat{v}} \, dv, \tag{3.16}$$

where $f_{Transformer}, g_{Transformer}$ represent the transformer encoding and decoding functions., respectively, as learned by the transformer-based variational autoencoder. The identifiability result of the TV-CANM holds under the following conditions: (i) the data are generated according to the cascade additive noise model, (ii) all noise variables are mutually independent, (iii) the structural functions are nonlinear and measurable, (iv) the transformer encoder consistently approximates the nonlinear causal mechanism, and (v) the causal direction is determined by comparing the test log-likelihood in both directions. Under these assumptions, the direction yielding the larger test log-likelihood corresponds to the true causal direction. The detailed derivation is shown in the supplementary file (Appendix).

4. Practical algorithm for the TV-CANM

The detailed steps of our proposed model approach are outlined in the algorithm. This algorithm guides the model through the learning and application phases by integrating the transformer with variational inference to determine causal direction.

Table 1. Detailed procedure of the TV-CANM-based causal direction inference algorithm.

Input: Dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$
Output: Inferred causal direction: $X \to Y$ or $Y \to X$

Step 1: Data normalization and splitting
(1) Normalize each variable to zero mean and unit variance.
(2) Create two directional datasets: $\mathcal{D}_{X\to Y} = (X, Y)$ and $\mathcal{D}_{Y\to X} = (Y, X)$.
(3) Split both datasets into training and testing sets.

Step 2: Model training
For each direction $d \in \{X \to Y,\ Y \to X\}$:
Initialize the TV-CANM model with a transformer encoder and decoder.
For each epoch $t = 1, \ldots, T$:
Encode (x, y) using the input projection and transformer encoder.
Compute μ and $\log(\sigma^2)$ from the encoded representation.
Sample latent variable $v = \mu + \exp(0.5 \log(\sigma^2))\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.
Decode $\hat{y} = f_{\text{Decoder}}([x, v])$.
Compute loss

$$\mathcal{L} = -\sum \log \mathcal{N}(y - \hat{y}; 0, sdy^2) + \beta\, \mathrm{KL}\left(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, 1)\right).$$

Update model parameters using backpropagation.
Compute average test loss, MSE, and RMSE.

Step 3: Causal direction selection
If Test Likelihood$(X \to Y) > (Y \to X)$, then
infer causal direction $(X \to Y)$; otherwise
infer $(Y \to X)$.

The following algorithm, known as causal direction inference using the TV-CANM, determines the causal relationship between two variables by employing a transformer-based variational autoencoder on datasets (X, Y) and (Y, X). Initially, the input pairs are normalized and divided into training and testing sets for each direction. A transformer encoder extracts latent features from the inputs, yielding the mean and log-variance of the latent variable, which are then sampled using the reparameterization trick. A fully connected decoder reconstructs the target variable from the sampled latent and the input. The model's training objective is to minimize a loss function that includes a reconstruction term, with a fixed standard deviation, and a KL -divergence regularizer.

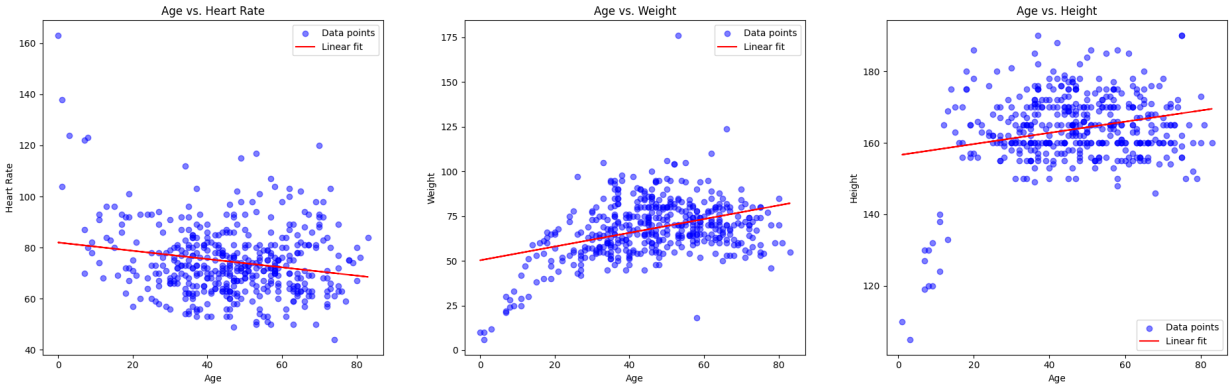
After training, the test log-likelihood of both directions is compared, and the direction that has a greater (less negative) likelihood is selected as the causal direction. Although MSE and RMSE are also

used to measure the model performance, the direction of causality is only established through the test likelihood.

5. Experiments

5.1. Dataset and setup

In the paper, the Arrhythmia dataset used was the publicly available dataset gathered by the author (see Figure 5) in the arrhythmia data (452 patient records with diverse physiological indicators of cardiac arrhythmia) (arrhythmia data, [42]). To narrow down on the age factor as a cause of physiological changes, the analysis only examined the variables of age, heart rate, weight, and height to isolate the known cause-and-effect associations. Age as an element that is inherently connected to the biological aging process affects height, weight, and heart rate. This choice aligns with the real processes that typically accompany aging, such as metabolic and cardiovascular changes. During preprocessing, two height observations (680 cm and 780 cm) were identified as physiologically impossible values and treated as recording errors. These observations were removed prior to model training to avoid numerical instability and distortion of parameter estimation. The remaining samples were processed identically for all experiments, ensuring that preprocessing did not alter the comparative evaluation of causal directions.



(a) Age versus Heart Rate (b) Age versus Weight (c) Age versus Height

Figure 5. Scatter plots showing the relationships between Age and each of Heart Rate, Height, and Weight, with fitted linear regression lines to illustrate trends across the three datasets.

Although the experimental evaluation is conducted on selected variable pairs for interpretability, the proposed TV-CANM architecture is designed for nonlinear high-dimensional data through transformer-based latent representation learning. The Arrhythmia dataset serves as a practical benchmark to demonstrate the applicability of the proposed framework. All competing methods were evaluated using the same data preprocessing, training/testing partition, and evaluation metrics to ensure a fair comparison.

5.2. Evaluation metrics

Commonly used evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and log-likelihood are widely adopted to evaluate and infer causal relationships [43, 44]. Therefore, this study employs these three metrics to assess model performance.

- Mean Squared Error (MSE): MSE measures the average squared difference between the actual values y_i and predicted values $\hat{y}_i$. The formula is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5.1}$$

- Root Mean Squared Error (RMSE): RMSE represents the square root of the average squared deviations between actual and predicted values, providing an error metric in the same units as the data. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{5.2}$$

- Log-Likelihood: The Gaussian log-likelihood evaluates how well the predicted values $\hat{y}_i$ align with the actual values y_i, assuming normally distributed errors. It offers a probabilistic measure of model fit by incorporating the error variance σ^2. The log-likelihood is computed as:

$$\log \mathcal{L} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{5.3}$$

  where $\sigma = \max(\text{sdy}, 0.05)$ denotes the clamped standard deviation (ensuring a minimum value of 0.05 for numerical stability.)

5.3. Causal analysis with existing models

We report the causal inference results obtained for the Arrhythmia dataset in Table 2 using some of the most common methods discussed: ANM with hilbert–schmidt independence criterion (HSIC) [45] and regression with XGBoost [46], IGCI [47], CAM [48], and PNL [23] as well as loss-based causal inference [49]. The goal is to identify the direction of the causal effect of age on each of the target variables heart rate, weight and height, as well as to validate whether the causal relationship is consistent with the domain expectations.

For all three data sets, the independence scores were lower with ANM (HSIC) than with rescaled ANM he Information-theoretic/Independent Score (ISIC), indicating reduced residual dependence. In all cases, IGCI consistently pointed in the direction of Age → Target, with negative scores supporting causal asymmetry. The results showed that a lower average loss for the Age → Target was achieved for all outcomes with CAM. Age was also identified as the causal factor by both PNL and loss-based methods, as indicated by a generally lower reconstruction loss in that direction.

Age was chosen because its causal effect on physiological characteristics is known in biomedical publications. The variable pairs chosen are biologically plausible examples for assessing the direction of causality in the identification process as proposed.

Table 2. Causal directions inferred by the existing model for each dataset, demonstrating its ability to identify the underlying cause-and-effect relationships across diverse data domains.

Models	Heart Rate Data Age(X)→Heart(Y) Heart(Y)→Age(X)		Weight Data Age(X)→Heart(Y) Heart(Y)→Age(X)		Height Data Age(X)→Heart(Y) Heart(Y)→Age(X)		Causal Direction	Ground Truth
	(X→Y)	(Y→X)	(X→Y)	(Y→X)	(X→Y)	(Y→X)		
ANM (HSIC)	0.0021	0.0024	0.0029	0.0032	0.0056	0.0061	Age→Heart Rate Age→Weight Age→Height	Age→Heart Rate Age→Weight Age→Height
IGCI (SCORE)	-45.7299	16.5797	-45.7299	29.5378	-28.3841	16.5797	Age→Heart Rate Age→Weight Age→Height	Age→Heart Rate Age→Weight Age→Height
CAM (Avg. Loss)	0.6654	0.7731	0.5593	0.6103	0.4864	0.7647	Age→Heart Rate Age→Weight Age→Height	Age→Heart Rate Age→Weight Age→Height
PNL	0.8988	0.9522	0.6826	0.7986	0.5639	0.8103	Age→Heart Rate Age→Weight Age→Height	Age→Heart Rate Age→Weight Age→Height
Loss -Based Causal Inference	0.8476	0.9540	0.6814	0.7425	0.8370	0.8642	Age→Heart Rate Age→Weight Age→Height	Age→Heart Rate Age→Weight Age→Height

5.4. Causal analysis with the TV-CANM

Each pairing (age with heart rate, weight, height) was tested in two models (CANM and TV-CANM) over three epochs or training periods (300, 400 , and 500). The task for each model was to find the right causal effect and test the validity of its causal inference at each epoch. Interestingly, both models reached the same results as the ground truth: Age → Heart Rate, Age → Height and Age , → Weight. On the other hand, the TV-CANM model failed to achieve lower loss values at each epoch compared to the CANM, suggesting that the TV-CANM is more robust in learning causal relationships and achieving higher accuracy and efficiency.

In this study, the causal relationship between age and some key physiological factors, including heart rate, weight, and height, was investigated using three datasets (within the Arrhythmia dataset). Age relationships with these variables are a biological prerequisite for causal inference studies because aging is associated with biological changes that require an understanding of their variation. The TV-CANM and CANM models correctly identified the right causal effects for all analyzed datasets (see Tables 3–5). For the Age and Heart Rate dataset (Table 3), both models correctly identified Age as the cause affecting Heart Rate. In Age and Height (Table 4) and Age and Weight (Table 5), both models were able to correctly find the correct relationships between Age → Height and Age → Weight, respectively.

Two assessment metrics, MSE and RMSE, were used to evaluate models under three training conditions at 300, 400, and 500 epochs. The results across all scenarios showed that the TV-CANM produced more accurate results than the CANM. Across all datasets, test likelihood remained consistent as an approach for establishing the correct causal direction. The results highlight the strong operational and effectiveness qualities of the TV-CANM, which stem from its encoder decoder design and its

self-attention capacity to track complex relationships in the data.

Table 3. The table presents the causal inference results for the Heart Rate dataset, where Age is considered the causal variable (X) and Heart Rate is the effect variable (Y).

Model	Epoch	Direction	Test Likelihood	MSE	RMSE	Suggested Causal Direction	Ground Truth
CANM	300	X→Y	-1.4742	1.1053	1.0513	X→Y	Age (X)→Heart Rate (Y)
		Y→X	-1.4853	1.1315	1.0637		
	400	X→Y	-1.4528	1.0289	1.0143	X→Y	Age (X)→Heart Rate (Y)
		Y→X	-1.4811	1.0861	1.0421		
	500	X→Y	-1.4665	1.0914	1.0447	X→Y	Age (X)→Heart Rate (Y)
		Y→X	-1.48622	1.1318	1.0638		
TV-CANM (Proposed)	300	X→Y	-1.4736	1.0992	1.0484	X→Y	Age (X)→Heart Rate (Y)
		Y→X	-1.4842	1.1244	1.0604		
	400	X→Y	-1.4548	1.0617	1.0304	X→Y	Age (X)→Heart Rate (Y)
		Y→X	-1.4665	1.0914	1.0447		
	500	X→Y	-1.4456	1.0132	1.0066	X→Y	Age (X)→Heart Rate (Y)
		Y→X	-1.4733	1.1061	1.0517		

Table 4. The table presents the causal inference results for the Height dataset, where Age is considered the causal variable (X) and Height is the effect variable (Y).

Model	Epoch	Direction	Test Likelihood	MSE	RMSE	Suggested Causal Direction	Ground Truth
CANM	300	X→Y	-1.3446	0.8505	0.9222	X→Y	Age (X)→Height (Y)
		Y→X	-1.3741	0.9086	0.9532		
	400	X→Y	-1.3286	0.8117	0.9009	X→Y	Age (X)→Height (Y)
		Y→X	-1.3752	0.9084	0.9531		
	500	X→Y	-1.3361	0.8234	0.9074	X→Y	Age (X)→Height (Y)
		Y→X	-1.3628	0.9053	0.9514		
TV-CANM (Proposed)	300	X→Y	-1.3619	0.8398	0.9164	X→Y	Age (X)→Height (Y)
		Y→X	-1.3724	0.9048	0.9512		
	400	X→Y	-1.3254	0.7099	0.8425	X→Y	Age (X)→Height (Y)
		Y→X	-1.3693	0.8963	0.9467		
	500	X→Y	-1.2755	0.6278	0.7923	X→Y	Age (X)→Height (Y)
		Y→X	-1.3723	0.8869	0.9417		

Table 5. The table presents the causal inference results for the Weight dataset, where Age is considered the causal variable (X) and Weight is the effect variable (Y).

Model	Epoch	Direction	Test Likelihood	MSE	RMSE	Suggested Causal Direction	Ground Truth
CANM	300	X→Y	-1.1819	0.5253	0.7248	X→Y	Age (X)→Weight (Y)
		Y→X	-1.2887	0.7386	0.8594		
	400	X→Y	-1.1820	0.5260	0.7253	X→Y	Age (X)→Weight (Y)
		Y→X	-1.2886	0.7391	0.8597		
	500	X→Y	-1.1815	0.5244	0.7241	X→Y	Age (X)→Weight (Y)
		Y→X	-1.2903	0.7579	0.8706		
TV-CANM (Proposed)	300	X→Y	-1.1757	0.5128	0.7161	X→Y	Age (X)→Weight (Y)
		Y→X	-1.2880	0.7286	0.8536		
	400	X→Y	-1.1783	0.5178	0.7196	X→Y	Age (X)→Weight (Y)
		Y→X	-1.2897	0.7386	0.8594		
	500	X→Y	-1.1794	0.5199	0.7210	X→Y	Age (X)→Weight (Y)
		Y→X	-1.2997	0.7412	0.8609		

6. Conclusions

This paper introduced a transformer-based variational autoencoder combined with the cascade additive noise model (TV-CANM) to enhance the detection of causal directions in nonlinear, high-dimensional datasets. Using the self-attention mechanism in transformers, the model can effectively capture both local and global relationships, improving performance over the standard CANM model. Mean squared error (MSE) and root mean squared error (RMSE) were used to assess the model's effectiveness, whereas causal directions were estimated using test likelihood.The TV-CANM generated lower error rates and precisely recognized causal directions aligned with domain knowledge. To establish the strength of our method, we compared the TV-CANM with established causal inference methods, including ANM, IGCI, CAM, PNL, and a loss-based method. These models are based on theoretical ideas such as noise independence (ANM, PNL) or independence between the input distribution and the causal mechanism (IGCI), all of which confirm real causal directions and support the validity of our suggested model. Future studies will focus on building an extended TV-CANM framework that includes a confounding cascade mechanism and accounts for unobserved intermediate variables and confounding structures. This expansion would enable more precise modeling of indirect effects and

hidden confounders, which are common in real-world data. Further, theoretical studies will investigate the identifiability of causal direction in the TV-CANM framework, especially in cases where the test likelihood can correctly identify the causal direction. These new developments will further enhance the model's interpretability and applicability for biomedicine [50], time-series analysis [51], and large-scale observation studies [52].

The Arrhythmia dataset has some clinical parameters that can affect the age/physiological parameter relationship. We used age as the main causal variable to evaluate the proposed framework in the present study. When using the framework to conduct a more extensive clinical study, other clinical factors like sex, drug use, disease severity, etc. can further clarify the causal interpretation of the findings.

Author contributions

Sohail Ahmad: Conceptualization, methodology, formal analysis, software, writing – original draft; Mona F. ElWakeel: Conceptualization, investigation, data curation, validation, funding acquisition, writing – review and editing; Moiz Qureshi: Methodology, software, formal analysis, visualization, writing – original draft; Hasnain Iftikhar: Supervision, Data curation, investigation, resources, validation, writing – review and editing; Paulo Canas Rodrigues: Supervision, project administration, funding acquisition, conceptualization, writing – review and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the General Directorate of Scientific Research & Innovation, Dar Al Uloom University, through the Scientific Publishing Funding Program.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Y. Li, O. Deng, A. Ogihara, S. Nishimura, Q. Jin, Causal discovery of health features from wearable device and traditional Chinese medicine diagnosis data, In: *HCI international 2023 – Late breaking papers*, 2023, 556–569. https://doi.org/10.1007/978-3-031-48041-6_37
2. C. Glymour, K. Zhang, P. Spirtes, Review of causal discovery methods based on graphical models, *Front. Genet.*, **10** (2019), 524. https://doi.org/10.3389/fgene.2019.00524
3. M. Huber, An introduction to causal discovery, *Swiss J. Econ. Stat.*, **160** (2024), 14. https://doi.org/10.1186/s41937-024-00131-4

4. L. Jin, Y. Zhang, Y. Li, S. Wang, H. H. Yang, J. Wu, M. Zhang, MoE^2: Optimizing collaborative inference for edge large language models, *IEEE/ACM Trans. Netw.*, **34** (2026), 4637–4651. https://doi.org/10.1109/TON.2026.3677371

5. H. Zhang, Y. Ren, Y. Xia, S. Zhou, J. Guan, Towards effective causal partitioning by edge cutting of adjoint graph, *IEEE Trans. Pattern Anal. Mach. Intell.*, **46** (2024), 10259–10271. https://doi.org/10.1109/TPAMI.2024.3435503

6. J. Pearl, *Causality*, 2Eds., Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9780511803161

7. M. Caliendo, S. Kopeinig, Some practical guidance for the implementation of propensity score matching, *J. Econ. Surv.*, **22** (2008), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

8. M. Baiocchi, J. Cheng, D. S. Small, Instrumental variable methods for causal inference, *Stat. Med.*, **33** (2014), 2297–2340. https://doi.org/10.1002/sim.6128

9. J. Bor, E. Moscoe, T. Bärnighausen, Three approaches to causal inference in regression discontinuity designs, *Epidemiology*, **26** (2015), e28–e30. https://doi.org/10.1097/EDE.0000000000000256

10. F. F. Stanaway, A. Diaz, R. Maddox, Causal inference, mediation analysis and racial inequities, *Int. J. Epidemiol.*, **53** (2024), dyae038. https://doi.org/10.1093/ije/dyae038

11. Q. Liu, P. Chen, K. Lin, K. Zhao, J. Ding, Y. Li, Sample-efficient backtrack temporal difference deep reinforcement learning, *Knowl.-Based Syst.*, **330** (2025), 114613. https://doi.org/10.1016/j.knosys.2025.114613

12. R. Visontay, L. Mewton, T. Slade, I. M. Aris, M. Sunderland, Moderate alcohol consumption and depression: A marginal structural model approach promoting causal inference, *Am. J. Psychiatry*, **180** (2023), 209–217. https://doi.org/10.1176/appi.ajp.22010043

13. Y. Lu, Q. Zheng, D. Quinn, Introducing causal inference using Bayesian networks and do-calculus, *J. Stat. Data Sci. Educ.*, **31** (2023), 3–17. https://doi.org/10.1080/26939169.2022.2128118

14. F. Podestà, Combining process tracing and synthetic control method: Bridging two ways of making causal inference in evaluation research, *Evaluation*, **29** (2023), 50–66. https://doi.org/10.1177/13563890221139511

15. B. Vakulenko-Lagun, C. Magdamo, M. L. Charpignon, B. Zheng, M. W. Albers, S. Das, causalCmprsk: An R package for nonparametric and Cox-based estimation of average treatment effects in competing risks data, *Comput. Methods Programs Biomed.*, **242** (2023), 107819. https://doi.org/10.1016/j.cmpb.2023.107819

16. Z. Yan, Z. Pan, G. Hu, H. Yan, Finite-time annular domain guaranteed cost control for uncertain mean-field stochastic systems with Wiener and Poisson noises, *IEEE Trans. Syst. Man Cybern. Syst.*, **55** (2025), 7981–7993. https://doi.org/10.1109/TSMC.2025.3598993

17. P. Spirtes, C. Glymour, R. Scheines, *Causation, prediction, and search*, MIT Press, 2001. https://doi.org/10.7551/mitpress/1754.001.0001

18. B. Huang, K. Zhang, J. Zhang, J. Ramsey, C. Glymour, B. Schölkopf, Causal discovery from heterogeneous/nonstationary data, *J. Mach. Learn. Res.*, **21** (2020), 1–53.

19. H. Huang, W. Zhang, Y. Xu, H. Zhang, J. Zou, A method for PWM frequency harmonic suppression using zero-vector switching SVPWM strategy in three-level inverter drive systems, *IEEE J. Emerg. Sel. Top. Power Electron.*, **13** (2025), 1482–1491. https://doi.org/10.1109/JESTPE.2024.3485962
20. K. Zhang, Z. Wang, J. Zhang, B. Schölkopf, On estimation of functional causal models: General results and application to the post-nonlinear causal model, *ACM Trans. Intell. Syst. Technol.*, **7** (2015), 13. https://doi.org/10.1145/2700476
21. P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf, Nonlinear causal discovery with additive noise models, *Adv. Neural Inf. Process. Syst.*, 2008, 689–696.
22. S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, A linear non-Gaussian acyclic model for causal discovery, *J. Mach. Learn. Res.*, **7** (2006), 2003–2030.
23. K. Zhang, A. Hyvarinen, On the identifiability of the post-nonlinear causal model, *arXiv preprint*, 2012, arXiv:1205.2599. https://doi.org/10.48550/arXiv.1205.2599
24. Z. Liang, L. Liu, J. Yang, J. Yu, S. Zhang, X. Zhang, et al., AUV cooperative localization method based on M–H resampling and state covariance matrix optimized particle filter, *IEEE Trans. Instrum. Meas.*, **74** (2025), 8513314. https://doi.org/10.1109/TIM.2025.3609326
25. X. Zheng, B. Aragam, P. K. Ravikumar, E. P. Xing, DAGs with no tears: Continuous optimization for structure learning, *Adv. Neural Inf. Process. Syst.*, **31** (2018).
26. Y. Yu, J. Chen, T. Gao, M. Yu, DAG-GNN: DAG structure learning with graph neural networks, *Proc. 36th Int. Conf. Mach. Learn.*, **97** (2019), 7154–7163.
27. P. Lippe, T. Cohen, E. Gavves, Efficient neural causal discovery without acyclicity constraints, *arXiv preprint*, 2021, arXiv:2107.10483. https://doi.org/10.48550/arXiv.2107.10483
28. I. Tsamardinos, L. E. Brown, C. F. Aliferis, The Max-min hill-climbing Bayesian network structure learning algorithm, *Mach. Learn.*, **65** (2006), 31–78. https://doi.org/10.1007/s10994-006-6889-7
29. J. M. Mooij, S. Magliacane, T. Claassen, Joint causal inference from multiple contexts, *J. Mach. Learn. Res.*, **21** (2020), 1–108.
30. Z. Zhou, X. Zhang, L. Zhang, J. Liu, E. Cambria, C. Li, FineFake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection, *Inf. Fusion*, **132** (2026), 104253. https://doi.org/10.1016/j.inffus.2026.104253
31. D. P. Kingma, M. Welling, Auto-encoding variational Bayes, *arXiv preprint*, 2013, arXiv:1312.6114. https://doi.org/10.48550/arXiv.1312.6114
32. A. Almodóvar, J. Parras, S. Zazo, Propensity weighted federated learning for treatment effect estimation in distributed imbalanced environments, *Comput. Biol. Med.*, **178** (2024), 108779. https://doi.org/10.1016/j.compbiomed.2024.108779
33. L. Wu, K. Wang, K. Nie, S. Guo, C. Gao, Z. Wang, S. Li, TFGIN: Tight-fitting graph inference network for table-based fact verification, *ACM Trans. Inf. Syst.*, **43** (2025), 130. https://doi.org/10.1145/3734520
34. S. Yin, Z. Xiang, Adaptive operator selection with dueling deep Q-network for evolutionary multi-objective optimization, *Neurocomputing*, **581** (2024), 127491. https://doi.org/10.1016/j.neucom.2024.127491

35. J. Qiao, R. Cai, K. Zhang, Z. Zhang, Z. Hao, Causal discovery with confounding cascade nonlinear additive noise models, *ACM Trans. Intell. Syst. Technol.*, **12** (2021), 1–28. https://doi.org/10.1145/3482879

36. S. Ahmad, H. Wang, Transformer-variational autoencoder for estimating individual treatment effect using causal inference framework, *Appl. Intell.*, **55** (2025), 855. https://doi.org/10.1007/s10489-025-06738-1

37. S. Ahmad, K. Shah, A. Debbouche, Structural equation modeling for causal effect estimation with machine learning, *J. Comput. Appl. Math.*, **475** (2026), 117020. https://doi.org/10.1016/j.cam.2025.117020

38. S. Ahmad, H. Wang, TV-CCANM: A transformer variational inference in confounding cascade additive noise model for causal effect estimation, *J. Stat. Comput. Simul.*, **95** (2025), 3048–3076. https://doi.org/10.1080/00949655.2025.2516793

39. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, *Adv. Neural Inf. Process. Syst.*, **30** (2017). https://doi.org/10.48550/arXiv.1706.03762

40. H. Qiu, J. Zhang, J. Zhou, G. Shi, G. Yao, X. Cai, A novel delta-type transformerless unified power flow controller with multiple power regulation ports, *IEEE Trans. Power Electron.*, **40** (2025), 15820–15834. https://doi.org/10.1109/TPEL.2025.3575074

41. X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, *Proc. Fourteenth Int. Conf. Artif. Intell. Stat.*, **15** (2011), 315–323.

42. J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf, Distinguishing cause from effect using observational data: Methods and benchmarks, *J. Mach. Learn. Res.*, **17** (2016), 1–102.

43. F. Quispe, E. Salcedo, H. Iftikhar, A. Zafar, M. Khan, J. E. Turpo-Chaparro, et al., Multi-step ahead ozone level forecasting using a component-based technique: A case study in Lima, Peru, *AIMS Environ. Sci.*, **11** (2024), 401–425. https://doi.org/10.3934/environsci.2024020

44. H. Iftikhar, M. Qureshi, J. Zywiołek, J. L. López-Gonzales, O. Albalawi, Short-term PM 2.5 forecasting using a unique ensemble technique for proactive environmental management initiatives, *Front. Environ. Sci.*, **12** (2024), 1442644. https://doi.org/10.3389/fenvs.2024.1442644

45. A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, A. Smola, A kernel statistical test of independence, *Adv. Neural Inf. Process. Syst.*, **20** (2007).

46. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, 785–794. https://doi.org/10.1145/2939672.2939785

47. D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, B. Schölkopf, Information-geometric approach to inferring causal directions, *Artif. Intell.*, **182–183** (2012), 1–31. https://doi.org/10.1016/j.artint.2012.01.002

48. P. Buhlmann, J. Peters, J. Ernest, CAM: Causal additive models, high-dimensional order search and penalized regression, *Ann. Stat.*, **42** (2014), 2526–2556. https://doi.org/10.1214/14-AOS1260

49. J. Mooij, D. Janzing, J. Peters, B. Schölkopf, Regression by dependence minimization and its application to causal inference in additive noise models, *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, 745–752. https://doi.org/10.1145/1553374.1553470

50. Y. Zhang, J. Xiang, J. Li, Denoising self-supervised learning for disease-gene association prediction, *BMC Bioinform.*, **26** (2025), 260. https://doi.org/10.1186/s12859-025-06281-3

51. H. Iftikhar, M. Khan, J. E. Turpo-Chaparro, P. C. Rodrigues, J. L. López-Gonzales, Forecasting stock prices using a novel filtering-combination technique: Application to the Pakistan stock exchange, *AIMS Mathematics*, **9** (2024), 3264–3288. https://doi.org/10.3934/math.2024159

52. Y. Tian, H. Tian, A multi-layer dynamic model of information propagation considering individual three-phase linear modulated behaviors, *Physica A*, **689** (2026), 131427. https://doi.org/10.1016/j.physa.2026.131427

AIMS Press